MANAV RACHNA INTERNATIONAL UNIVERSITY

Short Synopsis

For

Ph. D. Programme 2011-12

Searchable Encryption and Data Retrieval in cloud computing

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

FACULTY OF ENGINEERING & TECHNOLOGY

Submitted by:

Vasudha Arora
Registration No.:11/Ph. D/0014

Supervisor :

Dr. S.S. Tyagi
Professor & Head
Department of Computer Science & Engineering
Faculty of Engineering & Technology
Manav Rachna International University
Faridabad, Haryana

ABSTRACT

The driving motivation behind the cloud computing is to provide IT resources as a service that encapsulates other IT resources. Data service outsourcing is one of the service that is economically enabled by the cloud computing. The main threat on data privacy roots in the cloud itself. When users outsource their private data onto the cloud, the cloud service providers are able to control and monitor the data and the communication between users and the cloud. To ensure privacy, users usually encrypt the data before outsourcing it onto cloud, which brings great challenges to effective data utilization. However, even if the encrypted data utilization is possible, users still need to communicate with the cloud and allow the cloud to operate on the encrypted data, which potentially causes leakage of sensitive information. Although secure storage has already captured the attention of many cloud providers, offering a higher level of protection for their customer's data, we think that more advanced techniques such as searchable encryption and secure outsourced computation has become popular, opening the doors of the Cloud to the customers with higher security requirements. A concern of sensitive information on cloud potentially causes privacy problems. Data encryption protects data security to some extent, but at the cost of compromised efficiency. Searchable symmetric encryption (SSE) allows retrieval of encrypted data over cloud. These searchable encryption schemes allow retrieval of encrypted data over the cloud thereby making data retrieval fast and efficient.

These schemes guarantee the data security and preserve data privacy. During the whole process, no plain text data or keyword is visible to the cloud servers. To improve feasibility and to save the expense in the cloud paradigm, it is preferred to get the retrieval result with the most relevant files that match users' interest instead of all the files, which indicates that the files should be ranked in the order of relevance by users' interest and only the files with the highest relevance are sent back to users. Our motive is to suggest a secure scheme where the data owner can store their data on the Public Cloud while having the access control completely in their hands. Users have the ability to search through their data in a ranked manner thus saving the overhead involved in reading through each file for relevance in an encrypted database.

Key words: Data outsourcing, Symmetric Searchable encryption, Ranked query, Relevance score

Contents

S. No.	Description	Page No.
1	Introduction	1-3
2	Literature Review	4-6
3	Description of Broad Area/Topic	7-8
4.	Problem Identification/Objective of the Study	9
5	Methodology to be adopted	10
6	Proposed Time Frame (Gantt Chart)	11
7	References	12-14

Introduction

"Cloud" is a collective term for a large number of developments and possibilities. It is not an invention, but more of a "practical innovation", combining several earlier inventions into something new and compelling. Cloud computing merges several already available technologies: high bandwidth networks, virtualization, Web 2.0 interactivity, time sharing, and browser interfaces.

Cloud Computing gets its name as a metaphor for the Internet. In essence, cloud computing is a construct that allows us to access applications that actually reside at a location other than our computer or other internet connected device. The beauty of cloud computing is that another company hosts your application (or suite of applications). This means that they handle the cost of servers, they manage the software updates and depending on your contract you pay for the service.

Cloud Computing is a popular phrase that is shorthand for applications that were developed to be rich Internet applications that run on the Internet (or "Cloud"). In the Cloud Computing paradigm, software that is traditionally installed on personal computers is shifted or extended to be accessible via the Internet. These "Cloud applications" or "Cloud apps" utilize massive data centers and powerful servers that host web applications and web services. They can be accessed by anyone with a suitable Internet connection and a standard web browser.

The term cloud computing was first used in an academic context by Prof. Kenneth K Chellapa, who described it in 1997 at the Informs Conference in Dallas as "a computing paradigm where the boundaries of computing will be determined by economic rationale rather than technical limits".

Cloud Computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This Cloud model promotes availability and is composed of five essential characteristics (On-demand self-service; Broad network access; Resource pooling; Rapid elasticity; Measured Service), three service models Cloud Software as a Service (SaaS), Cloud Platform as a Service (PaaS). Cloud Infrastructure as a Service (IaaS) and four deployment models (Private Cloud; Community Cloud; Public Cloud and Hybrid Cloud).

Services

The term services in the cloud computing is the concept of being able to use reusable, fine-grained components across a vendor's network. This is widely known as "as a service".

Offerings with as a service as a suffix includes:

- Low barriers to entry, making them available to small businesses

- Large scalability
- Multitenancy, which allows resources to be shared by many users.
- Device independence, which allows users to access the systems on different hardware.

Infrastructure as a service (Iaas): It provides capability to the consumers to hire infrastructure components such as servers, storage and network and enables them to deploy and run software including operating system and applications. It serves as a foundation for other two layers (SaaS and PaaS). The consumer hires these resources as a service based on needs and pays only for usage.

Platform as a service (PaaS): It is provided to the consumer to deploy consumer created or acquired applications on the cloud service provider's infrastructure. PaaS can be defined as application development environment provided as a service by the cloud provider. The consumers use the platforms that include editor, compiler, build and deploy capabilities to develop their applications. When consumers write their applications to run over the PaaS provider's software platform, elasticity and scalability is guaranteed transparently by the PaaS platform.

Software as a service (SaaS): It is a capability provided to the consumer to use provider's applications running in a cloud infrastructure. SaaS is at the top most layer of cloud computing stack which is directly used by the end user. In SaaS model applications such as Email, instant messaging etc are offered as a service by the cloud provider. The cloud provider will host and manage the required infrastructure and applications to support these services.

Cloud paradigm is an upcoming technology which faces number of challenges. The major issues include data privacy and loss of control of data owner on the sensitive data. The main threat on data privacy roots in the cloud itself. When users outsource their private data onto the cloud, the cloud service providers are able to control and monitor the data and the communication between users and the cloud. To ensure privacy, users usually encrypt the data before outsourcing it onto cloud, which brings great challenges to effective data utilization. However, even if the encrypted data utilization is possible, users still need to communicate with the cloud and allow the cloud operates on the encrypted data, which potentially causes leakage of sensitive information. Although secure storage has already captured the attention of many cloud providers, offering a higher level of protection for their customer's data, we think that more advanced techniques such as searchable encryption and secure outsourced computation has become popular, opening the doors of the Cloud to customers with higher security requirements. The data is encrypted at the client side before it is send to the servers. A concern of sensitive information on cloud potentially causes privacy problems. Data encryption protects data security to some extent, but at the cost of compromised efficiency. Searchable symmetric encryption (SSE) allows retrieval of encrypted data over cloud. Many schemes have been generated to make search on encrypted data possible. A number of algorithms based on symmetric

searchable encryption schemes have been proposed. Using these earlier SSE schemes enable users to retrieve the cipher text securely but they provide only Boolean keyword search, i.e., these schemes can only search for if a keyword exists in a file or not and do not consider to check for the relevance of the keyword with the file. When we apply these encryption schemes directly on outsourced data in cloud environment they prove to have two major drawbacks. First, for each search request, users have to incur huge post processing overhead as users do not have any knowledge of encrypted cloud data. Second, it causes unnecessary network traffic because server on each search request will search out and send all the files just based on presence or absence of keyword in the files which is undesirable for cloud's pay-per-use model. There are two different approaches for the problem of searching on encrypted data. First one is to build up an index that for each keyword W to be searched, the index lists the documents that contain W. An alternative is to perform a sequential scan without an index. The disadvantage of using an index is that the storing and updating the index has a substantial overhead but sequential scan is not efficient enough when the data size is very large. For, very large size databases a common technique to speed up the searching is to use a pre-computed inverted index that maps each keyword with the data file in which that keyword appears.

Literature Review

Mell and Grance (2009) discussed about the basics of cloud computing. The five essential characteristics, benefits, three service models and four deployment models of cloud are discussed. It also discusses about the risks of cloud computing which are elaborated to be the availability, privacy and legislation, data theft and loss. It also throws light on the basic building blocks of cloud such as network virtualization, storage virtualization, server virtualization, application virtualization, desktop virtualization and automation.

Subashini and Kavitha (2011) discussed various cloud security issues in service delivery models. They discussed about the complexity of security at various cloud layers. There are three service delivery models namely infrastructure as a service (IaaS), platform as a service (PaaS), software as a service (SaaS). In this paper a discussion about a survey by IEEE & CSA is discussed which shows that security is needed both to accelerate cloud adoption on a wide scale and to respond to regulatory drivers.

Authors analyzed different security issues related to nature of service delivery models of cloud computing. According to them, the main security issues related to the software as a service models are as follows: Data security, Network security, Data locality, Data integrity, Data segregation, Data Access, Authentication and authorization, Data Confidentiality, Data breaches, Virtualization vulnerability, backup of data. They explained data security as Security of client's personnel sensitive data on the cloud, so cloud provider must provide additional security features apart from default one that is used in traditional systems; it involves strong encryption techniques for data security. They described network security as securing communication over the internet as all the communication between cloud service provider and customer is done through internet, it involves the use of strong network traffic encryption techniques such as Transport layer security, secure socket layer.

Huang and Liao (2012) proposed a robust and searchable encryption scheme for data outsourcing in cloud computing by considering both numeric and non-numeric data. In the other field, this scheme provides some fault-tolerance availability for cloud computing. After all, we can protect the keyword search and data stored in the third party of cloud databases and storages from revealing to malicious adversaries.

Wang et al. (2012) In this paper authors discussed data privacy, sensitive cloud data have to be encrypted before outsourced to the commercial public cloud, which makes effective data utilization service a very challenging task. In this paper, authors define and solve the problem of secure ranked

keyword search over encrypted cloud data. Ranked search greatly enhances system usability by enabling search result relevance ranking instead of sending undifferentiated results, and further ensures the file retrieval accuracy. Specifically, they explored the statistical measure approach, i.e., relevance score, from information retrieval to build a secure searchable index, and developed a one-to-many order-preserving mapping technique to properly protect those sensitive score information.

Yu et al. (**2013)** Data encryption protects data security to some extent, but at the cost of compromised efficiency. Searchable symmetric encryption (SSE) allows retrieval of encrypted data over cloud. In this paper, authors proposed a new searchable encryption scheme where the data owner encrypts the searchable index with Homomorphic encryption. When the cloud server receives a query consisting of multi keywords , it computes the scores from the encrypted index stored on the cloud and then returns the encrypted scores of the files to the data user. Now, the data user decrypts the score and picks the top-k highest scoring files' identifiers to request the cloud server. The retrieval takes a two-round communication between the cloud server and the data user hence the name two round searchable encryption.
In TRSE, authors employed a vector space model and homomorphic encryption. The vector space model helps to provide sufficient search accuracy, and the homomorphic encryption enables users to involve in the ranking while the majority of computing work is done on the server side by operations only on ciphertext. As a result, information leakage can be eliminated and data security is ensured. Thorough security and performance analysis show that the proposed scheme guarantees high security and practical efficiency.

Xu et al. (2012) In this paper, authors aim to provide a viable solution for multikeyword ranked query problems over encrypted data in the cloud environment. They analyzed the existing solutions and design a novel algorithm called MKQE to address the issues. MKQE uses a partitioned matrices approach. When the amount of encrypted data increases and more keywords need to be introduced, the searching infrastructure can be naturally expanded with the minimal overhead. They also design a new trapdoor generation algorithm, which can solve the out-of-order problem in the returned result set without losing the data security and privacy property. Furthermore, the weights of the keywords are taken into consideration in the ranking algorithm when generating the query result. The DC (data consumer) has high probability to retrieve the files they really need.

Gupta et al. (2013) have suggested a secure scheme where the data owner can store their data on the Public Cloud while having the access control completely in their hands. User have the ability to search through their data in a ranked manner thus saving the overhead involved in reading through

each file for relevance in an encrypted database. The architecture follows a traditional Client / Server model with the client performing all the cryptographic operations and the server performing the search operations in a secure and confidential manner. Preliminary results show that security overhead in terms of space is minimal under this scheme

Mohamed et al. (2011) authors compared 8 symmetric key algorithms by testing them on 16 NIST statistical tests to select the highest security algorithms. Authors used eight modem encryption algorithms namely RC4, RC6, MARS, AES, DES, 3DES, Two-Fish, and Blowfish. The comparison was based on P-value and rejection rate. They also compared between eight encryption algorithms based on the speed of encryption/decryption to get the faster algorithm. As a conclusion AES was found to be fastest among all which takes least time to encrypt among all and is also the highest security algorithm.

Xiao and Yen (2012) In this paper, authors discussed the security of OPE schemes by analyzing the number of bits of the plaintext that remain secret from the adversary against a known plaintext attack with n known plaintexts The results presented in the paper not only help improve our understanding of the security of OPE schemes and guide its parameter selections, but also provide a general method for analyzing their security.

Description of Broad Area

Searchable Encryption and data retrieval in cloud computing:

Cloud computing has become one of the most significant issues in recent years. Its benefits have attracted almost every sector. However, in order to enjoy the wide utilization of cloud computing services through wired/wireless network, assurance of cloud security in terms of data confidentiality, authentication, non-repudiation and integrity is a critical factor for its success. Despite of all hype about the cloud, enterprise customers are still reluctant to deploy their business in the cloud. Security and loss of control are the major issues which reduces the growth of cloud computing.

Despite of its advantages, the cloud computing infrastructure faces very challenging tasks, especially on data privacy, security, and reliability issues. As data is placed on public clouds which are out of their trusted domains, data owners do not have direct control over their sensitive data and are worried about possible data loss and/or illegal use of their private data. Usually, cloud servers are considered as curious and untrusted, therefore data owners hesitate to adopt cloud technologies as there are risks of data exposure to the cloud service provider or a third party. Thus, providing sufficient security and privacy protections is very important, especially for applications processing health information, financial system and government data etc. To prevent information disclosure, the mainstream solution is to encrypt private data before uploading onto the cloud. On one hand, it ensures that data is not visible to external users and cloud administrators. On the other hand, there are severe processing limitations on encrypted data. For example, standard plain text based searching algorithms are not applicable any more. To perform a keyword based query, the entire data set has to be decrypted even if the matching result set is very small. It poses unbearable query latency and incurs unacceptable computational overhead. Current solutions use the following strategy to provide keyword based searching capabilities on encrypted data. An index vector is calculated for each individual file maintaining the information of keywords. Then, an index file which combines all index vectors is generated. The index file is also encrypted. Then, both the encrypted data and the encrypted index file are uploaded onto data center servers in the cloud. The cloud servers can then support cipher text based queries as follows, a data consumer submits a keyword based query, and the encrypted keywords are sent to the cloud server. The cloud server conducts a search on the encrypted index and returns a list of most relevant files. The user makes the decisions which files are needed and retrieve them from the server. After receiving encrypted files, the user decrypts the files

with the associated key. This approach can guarantee the data security and preserve data privacy. During the whole process, no plain text data or keyword is visible to the cloud servers.

To improve feasibility and save on the expense in the cloud paradigm, it is preferred to get the retrieval result with the most relevant files that match users' interest instead of all the files, which indicates that the files should be ranked in the order of relevance by users' interest and only the files with the highest relevance are sent back to users. Our motive is to suggest a secure scheme where the data owner can store their data on the Public Cloud while having the access control completely in their hands. User have the ability to search through their data in a ranked manner thus saving the overhead involved in reading through each file for relevance in an encrypted database.

Objectives

- To test the existing symmetric encryption algorithms on various NIST (National Institute for Standards and Technology) statistical tests to find the highest security symmetric key algorithm with least overhead.

- To develop and design a symmetric searchable encryption algorithm suitable for making a search over encrypted text as well as for numeric cloud data.

- To evaluate the efficiency of algorithm based on different parameters such as speed of encryption and decryption, time taken to generate a trapdoor, time taken to search.

- To analyze the efficiency of algorithm for relevance top-k retrieval over encrypted cloud data so as to reduce the cost in cloud pay per use model.

Methodology

Phase 1 Literature Review

Phase 2 Study of available tools, mechanisms, categorizing issues and challenges, evaluating disadvantages of existing systems and identifying areas of improvisation to make cloud adoption more comfortable to the users.

Phase 3 Formulating the research objectives- The Problem statement

Phase 4 Implementation-to achieve research objectives either through simulations or instances provided by cloud service providers.

Phase 5 Paper presentation and publications based on my research work.

Phase 6 Final report writing, reviewing of thesis and submission.

Proposed Time Frame (Research Plan)

Research Plan

	Sem 1 & sem 2 (June ’12 to Dec 2012)	Sem 3 (June ‘13)	Sem 4 (Dec ’13)	Sem 5 (june ’14)	Sem 6 (Dec ’14)	Sem 7 (June 15)	Sem 8 (Dec 15)	Sem 9 (June ‘16)
Course work								
Literature review								
Problem Formulation								
Implementation								
Testing and Analysis								
paper presentation and publication								
Thesis Submission								

References

Agrawal, R., Kiernan, J., Srikant, R., & Xu, Y., 2004 'Order Preserving Encryption for Numeric Data', SIGMOD '04 Proc. ACM SIGMOD International Conference on Management of Data, pp 563-574.

Agarwal, S., & Choubey, A., 2014, 'Survey of Fully Homomorphic Encryption and Its Potential to Cloud Computing Security', International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 7.

Akriti, P., Ann, M., & Sarvanan, D., 2014,'Ranked Keyword Search Using RSE over Outsourced Cloud Data', IPASJ International Journal of Computer Science, Vol. 2, Issue 3.

Bellare, M., Boldyreva, A., & O'Neill, A., 2009 'Deterministic and efficiently searchable encryption,' in proc. of Eurocrypt'09, vol. 5479 of LNCS, Springer.

Boldyreva, A., Chenette, N., Lee, Y., & Neill, A., 2011, ' Order Preserving Symmetric Encryption', Advances in Cryptology- CRYPTO 2011, 31st Annual International Cryptography Conference, LNCS, Springer.

Boneh, D., Crescenzo, G., Ostrovsky, R., & Persiano, G., 2004, 'Public Key Encryption with Keyword Search', Proc. International Conference on Theory and Applications of Cryptographic Techniques (Eurocrypt).

Buyya, R., Broberg, J., & Goscinski, A., 2011 "Cloud Computing principle and paradigm", wiley publishings.

Church, K. W., & Hanks, P., 1990, 'Word Association Norms, Mutual Information, and lexicography', Computational Linguistics, Vol. 16, pp 22-29.

Curtmola, R., Garay, J. A., Kamara, S., & Ostrovsky, R., 2006 'Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions,' Proc. ACM 13th Conf. Computer and Comm. Security (CCS).

Erl, T., Puttini, R., & Mahmood, Z.,'Cloud Computing: Concepts, Technology & Architecture', PHI kindle edition.

Girme, M. R., & Bhandari, G. M., 2013, 'Efficient Secure Ranked Search Algorithms over outsource Cloud Data', International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Vol. 2, Issue 5.

Goh, E. J., 2003, 'Secure indexes', Cryptology eprint Archieve, Report 2003/216, http://eprint.iacr.org/.

Gupta, S., Satapathy, S. R., Mehta, P., & Tripathy, A., 2013 'Secure and Searchable Data Storage in Cloud Computing', 3rd IEEE International Advanced Computing Conference (IACC), pp. 106-109.

Huang, J., & Liao, I., 2012, 'A searchable Encryption scheme for outsourcing cloud storage, COMMENTSTAT, pp. 142-146.

Joy, E. C., & Indira, K., 2014, 'Multi keyword Ranked Search over Encrypted Cloud Data', International Journal of Applied Engineering Research, Vol. 9, No. 20, pp 7149-7176.

Kandukuri, B. R., Ramakrishna, V., & Rakshit, A., 2009, 'Cloud Security Issues', IEEE International Conference on service computing (SSC 2009), pp. 517-520, Bangalore.

Mell, P., & Grance, T.,2009, 'The NIST Definition of Cloud Computing', Ver 15.

Mohamed, E. M., Abdelkader H.S., El-Etriby, S., 'Enhanced Data Security Model for Cloud Computing' , 8th Intenational Conference on Informatics & Systems (INFOS2012), cloud and mobile computing track, pp 12-17.

Pillai, L. K., & Saranya, P., 2014 'Towards Secure Top-K Multi Keyword Retrieval over Encrypted Cloud Data', International Journal of Scientific Research and Management (IJSRM), Vol. 2, Issue 6, pp 950-955.

Priya, P. S., Preethi, D., Priya, J., & Shanthini, B., 2014, 'Retrieval of Encrypted Data Using Multi Keyword Top-k Algorithm', International Journal of Scientific and Research Publications, Vol. 4, Issue 4.

Singhal, A., 2001, 'Modern Information Retrieval: A brief overview,' IEEE Data Engineering Bulletin, Vol. 24, No. 4, pp 35-43.

Song, D., Wanger, D., & Perrig, A., 2000, 'Practical Techniques for Searches on Encrypted Data', Proc. IEEE symp. Security and Privacy.

Subashini, S., & Kavitha, S. V., 2011, 'A survey on security issues in service delivery models of cloud computing', Journal of Network and Computer Applications Vol. 34, issue 1, pp.1-11.

Sun, X., Zhu, Y., Xia, Z., & Chen, L., 2014, 'Privacy-Preserving Keyword-based Semantic Search over Encrypted Cloud Data', International Journal of Security and Its Applications, Vol. 8, No. 3, pp 9-20.

Swaminathan, A., Mao, Y., Su, G. M., Gou, H., Varna, A. L., He, S., Wu, M., & Oard, D. W., 2007, 'Confidentiality-Preserving Rank-Ordered Search', Proc. Workshop Storage Security and Survivability.

Wang, C., Cao, N., Li, J., Ren, K., & Lou, W., 2010, 'Secure Ranked Keyword Search over Encrypted Cloud Data', proc. IEEE 30th International Conference on Distributed Computing Systems(ICDCS '10).

Wang, C., Cao, N., Ren, K., & Lou, W., 2012 'Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data,' IEEE Transactions On Parallel and Distributed Systems, Vol. 23, No. 8.

Xiao, L., & Yen, I., 2012, 'Security Analysis for Order Preserving Encryption Schemes', IEEE 46th annual conference on Information Sciences & Systems (CISS), pp 1-6.

Xu, Z., Kang, W., Li, R., Yow, K., & Xu, C., 2012 'Efficient multi-keyword ranked query on encrypted data in cloud', IEEE 18th International conference on parallel and distributed systems, pp. 239-249.

Yu, J., Lu, P., Zhu, Y., Xue, G., & Li, M., 2013 'Toward Secure Multikeyword Top-k Retrieval over Encrypted Cloud Data', IEEE Transactions on Dependable and Secure Computing, vol. 10, No. 4.

Zerr, O. S., Olmedilla, D., Nejdl, W., &. Siberski, W., 2009, 'Zerber+r: Top-k Retrieval from a Confidential Index', Proc. 12th International Conference on Extending Database Technology: Advances in Database Technology (EDBT).

Zittrower, S., & Zou, C. C., 2012, 'Encrypted Phrase Searching in the Cloud', IEEE Globecom 2012-Communication and Information System Security Symposium.